# Data Statistics Theory

Kristhian Santiago Palomino Fajardo

May 28, 2024

## 1 Introduction

In this document, we explain some fundamental statistical measures and provide practical examples for each one. The goal is to make these measures understandable for anyone, regardless of their previous training in statistics. Additionally, we introduce basic statistical data types to further aid in data analysis comprehension.

## 2 Data Types in Statistics

Statistical data can broadly be classified into two types: categorical and numerical, each serving different purposes in statistical analysis.

### 2.1 Categorical Data

Categorical data represent types of data which may be divided into groups. Examples of categorical data include:

- **Nominal data**: Data without any order (e.g., gender, race, color).
- **Ordinal data**: Data with a natural order but not evenly spaced (e.g., rankings, scale of satisfaction from poor to excellent).

### 2.2 Numerical Data

Numerical data are quantities and can be split into two subtypes:

- **Discrete data**: Countable data, such as the number of students in a classroom.
- **Continuous data**: Data that can take any value within a given range, such as height or weight.

## 3 Mean

The **mean** or average is the value obtained by summing all data points and dividing the result by the total number of data points.

### Example

Consider the values: 5, 3, 7, 3, and 8. The mean is calculated as:

$$\text{Mean} = \frac{5+3+7+3+8}{5} = 5.2$$

## 4 Median

The **median** is the value that lies in the middle of an ordered dataset. If there is an even number of observations, the median is the average of the two central values.

### Example

For the dataset: 1, 3, 3, 6, 7, 8, 9, the median is 6.

## 5 Trimmed Mean

The **trimmed mean** is a measure of central tendency that involves calculating the average after removing outliers.

### Example

Removing the top and bottom 20% of values in the set: 2, 2, 3, 5, 7, 9, 9, its trimmed mean would be:

$$\text{Trimmed Mean} = \frac{3+5+7}{3} = 5$$

# 6 Mean Absolute Deviation

The **mean absolute deviation** is the average of the absolute differences between each data point and the dataset's mean.

### Example

With the dataset: 2, 4, 6, 8, 10, the mean absolute deviation is calculated as follows:

$$\begin{aligned}
&\text{Mean} = 6, \\
&\text{MAD} = \frac{|2-6|+|4-6|+|6-6|+|8-6|+|10-6|}{5} \\
&\quad\;\; = 2.8
\end{aligned}$$

# 7 Standard Deviation

The **standard deviation** is a measure that indicates the amount of variation or dispersion of a dataset.

### Example

For the values: 2, 4, 4, 4, 5, 5, 7, 9:

$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}} = 2.138$$

# 8 Quantiles

**Quantiles** are values that divide the dataset into equal parts. Percentiles and quartiles are common examples of quantiles.

### Example

In an ordered dataset: 1, 2, 3, 4, 5, 6, 7, 8, 9, the first quartile (Q1) is 3 and the third quartile (Q3) is 7.